D586–D589 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj027

QTL MatchMaker: a multi-species quantitative trait loci (QTL) database and query system for annotation of genes and QTL

Kremena V. Star[1,2,*], Quingbin Song[2], Andy Zhu[2] and Erwin P. Böttinger[2,*]

[1]Albert Einstein College of Medicine and [2]Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1243, New York, NY 10029, USA

Received August 11, 2005; Revised and Accepted September 22, 2005

ABSTRACT

Identifying genes that underlie quantitative trait loci (QTL) is a challenging task. Here, we present a new QTL software system, named QTL MatchMaker. The system is designed to integrate and mine QTL information across human, mouse and rat genomes and to annotate functional genomic data. It combines and organizes information from relevant public databases and publications and integrates QTL, physical, genetic and cytogenetic maps across human, mouse and rat. To make this application available to the research community we have developed a website for high-throughput mapping of expressed sequences to QTL and for selection of candidate genes in the physiological genomics context of complex traits. QTL MatchMaker is accessible at http://pmrc.med.mssm.edu:9090/QTL/jsp/qtlhome.jsp

INTRODUCTION

A quantitative trait locus (QTL) is a polymorphic locus containing alleles that differentially affect the expression of a specific phenotypic trait (a genetic basis for physiological variation) (1). The purpose of QTL research is to identify genes and gene variants that underlie these traits. QTL are identified via association of the studied traits with genetic markers. The association strength is reported as a linkage score. Over the past 10 years QTL mapping has resulted in identification of thousands of chromosomal regions containing genes predicted to be involved in many complex human diseases, such as diabetes, hypertension, obesity and many others (2). Despite the accelerated rate of QTL discovery in human and mammalian models, finding the genes that underlie quantitative traits remains a challenge (3).

Several large-scale efforts collect and organize mammalian QTL and linkage groups: Mouse Genome Database (MGD) (4), Rat Genome Database (RGD) (5), Online Mendelian Inheritance in Man (6) and Entrez Gene (7). Currently, QTL are not fully integrated with the different types of chromosomal maps. For example, QTL mouse data are presented as genetic chromosomal positions while human linkage groups are reported as cytogenetic bands and rat QTL are positioned on radiation hybrid maps. In addition, automated, high-throughput tools to map expressed sequences and genes to QTL are not available to date. Here we present a multi-species QTL database and analysis software named QTL MatchMaker (http://pmrc.med.mssm.edu:9090/QTL/jsp/qtlhome.jsp). QTL MatchMaker is a web-based application that processes data from multiple sources and provides a web site for the analysis of candidate genes in the context of genetic traits. It accepts a list of GenBank accession numbers and mRNA RefSeq numbers for genes of interest. QTL MatchMaker then performs mapping of these genes against a curated QTL database and generates reports of co-localization events. A results file shows the genes positions and information about the corresponding QTL.

IMPLEMENTATION AND DATABASE STRUCTURE

The QTL MatchMaker database schema has been created in Oracle database server, which is running on a Linux platform. The user interface is written in Java programming language and communicates with the database through standard JDBC adapters.

Database design

To provide for automated mapping of experimentally-derived candidate genes to QTL, we compiled human, mouse and rat QTL, as well as human linkage groups information from PubMed articles, MGD, Entrez Gene and RGD websites.

*To whom correspondence should be addressed. Tel: +1 212 241 1884; Fax: +1 212 849 2643; Email: kvstar@aecom.yu.edu
Correspondence may also be addressed to Erwin P. Böttinger. Tel: +1 212 241 0800; Fax: +1 212 849 2643; Email: erwin.bottinger@mssm.edu

Table 1. QTL Database statistics

Species	Distinct traits	Total QTL/ linkage groups	Average interval length (bp)
Mus musculus	323	1475	39 864 556
Rattus norvegicus	81	629	40 498 069
Homo sapiens	109	143	24 336 058

A subset of human linkage groups was identified from Entrez Gene. Additional groups were collected from published reports on human linkage groups and QTL that were retrieved by text mining of PubMed abstracts with keywords: 'linkage group' and 'LOD score'. Since MGD and RGD maintain a list of mouse and rat QTL references, we used this information to download the original publications and extract the information described below to populate the QTL MatchMaker database.

QTL symbols, names, traits, flanking and peak markers associated with each QTL were collected from these publications along with LOD scores, parental strains, type of cross or epidemiologic data in cases of human linkage analysis. For QTL annotation we adopted Mammalian Phenotype ontology (5) developed at MGD and RGD. To assign base pair and cytogenetic positions to all flanking and peak markers, we downloaded the UCSC Genome Browser (8) 'STS' and 'cytoBand' positional tables and parsed them using the names of markers as keywords. This step is automated and allows for periodic updates of the marker positions.

To assign a search interval, QTL borders must be mapped experimentally via linkage crosses. However, a significant number of QTL are reported only by their peak markers. To address this gap we calculated the average QTL length in each species (mouse, rat and human) (Table 1). In cases where QTL flanking markers are not known, we assigned an interval by centering the average QTL length in the respective species on the QTL peak marker. The intervals generated in this way are flagged to mark whether the QTL of interest has experimentally or statistically defined borders.

To provide for cross-species comparison of QTL, we have downloaded the UCSC Genome Browser (8) 'hgBlastTab', 'mmBlastTab' and 'rnBlastTab' positional tables which allow for finding predicted orthologs using an accession number of interest as query. As a result the user can look directly at QTL that are positioned over a candidate gene in all three species automatically. These tables are also updated with every new release of human, mouse and rat genome assemblies.

All this information is curated and maintained in a central database implemented on Oracle server. In addition we have developed a standardized form for submitting new QTL information. This allows researchers to upload new QTL to the database making it a dynamic, up-to-date tool. As a result, QTL MatchMaker provides the users with the most current and accurate information about genes and QTL co-localization events.

Query and data retrieval

QTL information is retrieved via a web interface implemented in Java programming language. The QTL MatchMaker home page allows the user to choose a species of interest among mouse, human and rat. This choice dynamically leads to a second page where the traits of the chosen species are listed for further selection. The user is asked to enter a Genbank accession number or a mRNA RefSeq number for the gene of interest. We have downloaded the UCSC Genome Browser positional tables for human, mouse and rat known genes and expressed sequence tags to the QTL database. Based on these tables each accession number is assigned start and end base pair chromosomal positions for the corresponding transcript. Base pair position numbers of the transcript beginning (5′) and ending (3′) nucleotides are compared with the QTL base pair positions on the respective chromosome of the selected species. If a gene is located within a QTL, information for the QTL is linked with the gene data. To enable parallel searches for more than one gene we have developed a batch query function of the QTL MatchMaker. This allows for high-throughput mapping of genes of interest and QTL. In instances where a gene maps to multiple QTL, all matches are reported. The application generates a report file which presents the position of the genes and the QTL characteristics (Figure 1A). When analyzing this report file it is important to note that although a lot of genes can be in a QTL interval only a small percentage of those genes will be causal. The results can be downloaded in a Microsoft Excel or HTML file format.

From the QTL MatchMaker home page the user can also access the Cross Species search page and find predicted orthologs of the candidate gene. Orthologs are reported from the 'mmBlastTab', 'hgBlastTab' and 'rnBlastTab' UCSC Genome Browser positional tables which contain the Blastp results of known genes from the query and target species. All orthologs are assigned start and end base pair chromosomal positions and matched with the QTL in the respective species. The end report provides information on each QTL and the gene located inside it. As a result the user can identify QTL that are positioned over a gene of interest across all three species automatically (Figure 1B).

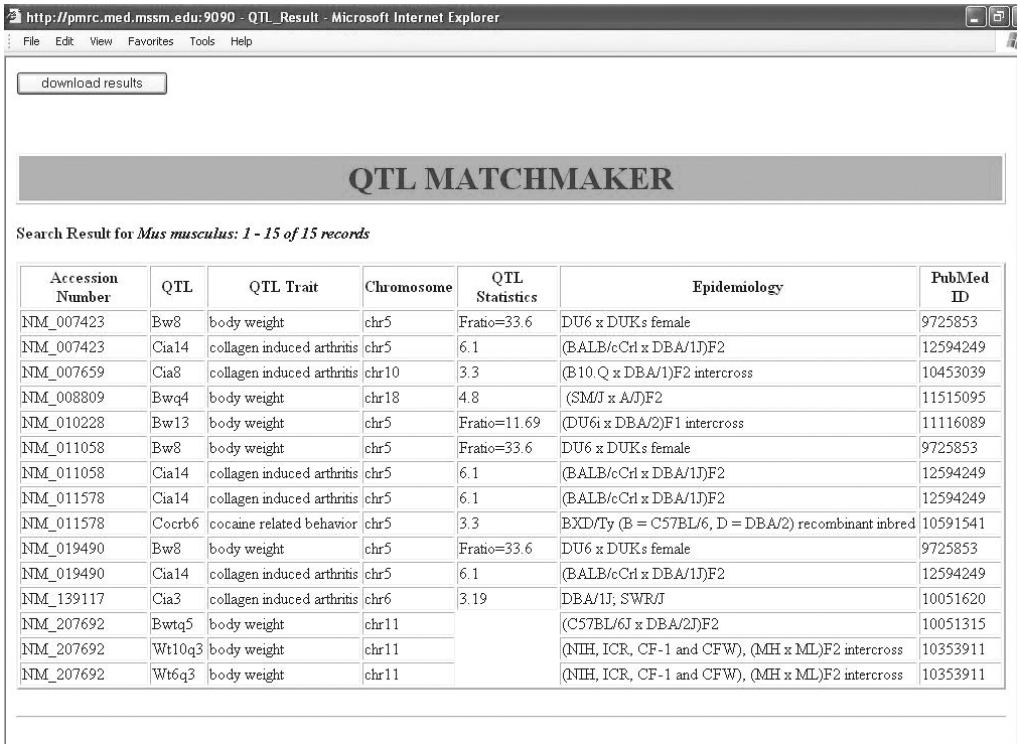
To address the difficulties in resolving QTL architecture we prioritize QTL candidate genes by applying a genetical genomics approach (9). We currently use the QTL MatchMaker for co-localizing transcripts culled from microarray experiments to physiologically related QTL in mouse, rat and human. The comparative genomics application of this tool allows mapping of experimentally-defined genesets from transcriptome analysis of human diseases to related QTL in mouse and rat. For example, we used the QTL MatchMaker to map mouse and rat orthologs of a human geneset identified by microarray analysis of kidney biopsy material from patients with systemic lupus erythematosus (10) to autoimmune QTL in rodents.

Another application which is currently being explored is the mapping of genes identified in mutagenesis screens to QTL. In cases when a mutated gene shows a phenotype similar to the trait characterizing the QTL it co-localizes to, this gene becomes a high priority candidate for resolving the QTL genetic structure.

DATA AVAILABILITY

All the data within the QTL MatchMaker database are freely available to the scientific community. In addition to the access through the user interface, the content of the database can be downloaded as Microsoft Excel files.

A



QTL MATCHMAKER

Search Result for *Mus musculus: 1 - 15 of 15 records*

Accession Number	QTL	QTL Trait	Chromosome	QTL Statistics	Epidemiology	PubMed ID
NM_007423	Bw8	body weight	chr5	Fratio=33.6	DU6 x DUKs female	9725853
NM_007423	Cia14	collagen induced arthritis	chr5	6.1	(BALB/cCrl x DBA/1J)F2	12594249
NM_007659	Cia8	collagen induced arthritis	chr10	3.3	(B10.Q x DBA/1)F2 intercross	10453039
NM_008809	Bwq4	body weight	chr18	4.8	(SM/J x A/J)F2	11515095
NM_010228	Bw13	body weight	chr5	Fratio=11.69	(DU6i x DBA/2)F1 intercross	11116089
NM_011058	Bw8	body weight	chr5	Fratio=33.6	DU6 x DUKs female	9725853
NM_011058	Cia14	collagen induced arthritis	chr5	6.1	(BALB/cCrl x DBA/1J)F2	12594249
NM_011578	Cia14	collagen induced arthritis	chr5	6.1	(BALB/cCrl x DBA/1J)F2	12594249
NM_011578	Cocrb6	cocaine related behavior	chr5	3.3	BXD/Ty (B = C57BL/6, D = DBA/2) recombinant inbred	10591541
NM_019490	Bw8	body weight	chr5	Fratio=33.6	DU6 x DUKs female	9725853
NM_019490	Cia14	collagen induced arthritis	chr5	6.1	(BALB/cCrl x DBA/1J)F2	12594249
NM_139117	Cia3	collagen induced arthritis	chr6	3.19	DBA/1J; SWR/J	10051620
NM_207692	Bwtq5	body weight	chr11		(C57BL/6J x DBA/2J)F2	10051315
NM_207692	Wt10q3	body weight	chr11		(NIH, ICR, CF-1 and CFW), (MH x ML)F2 intercross	10353911
NM_207692	Wt6q3	body weight	chr11		(NIH, ICR, CF-1 and CFW), (MH x ML)F2 intercross	10353911

B

QTL MATCHMAKER

Orthologs Report

Species	Accession Number	Transcription Start (bp)	Transcription End (bp)	QTL Symbol	Trait	Chromosome	QTL Statistics	Parents	PubMed ID#
Mm	AK035956	104389982	104512961	Axtofd4	anxiety-open field defecation	chrX	5.3	(H1 x L1)F2 intercross; (H2 x L2) F2 intercross	11454769
Mm	AK035956	104389982	104512961	Bw12	body weight	chrX	NONE	DU6 x DUKs female	9725853
Mm	AK035956	104389982	104512961	Gct4	granulosa cell tumorigenesis	chrX	chi square = 9.22; P<0.001	reciprocal F1 x F1 intercrosses [(SJL x SWR)F1 x (SJL x SWR) F1; (SJL x SWR)F1 x (SWR x SJL)F1; (SWR x SJL)F1 x (SJL x SWR)F1; (SWR x SJL)F1 x (SWR x SJL)F1]	9721880
Mm	AK035956	104389982	104512961	Lwq8	liver weight	chrX	NONE	DU6 x DUKs female	9725853; 11116089
Mm	AK035956	104389982	104512961	Mnic3	macronutrient intake, carbohydrate	chrX	4	F2 cross(C57BL/6J and CAST/EiJ)	12388789
Mm	AK035956	104389982	104512961	Mnif3	macronutrient intake, fat	chrX	4	F2 cross(C57BL/6J and CAST/EiJ)	12388789
Rn	L13722	101827289	102001374	Cia19	joint/bone inflammation maximum score	chrX	4.4	F2 BB/DRxBN	11001927
Rn	L13722	101827289	102001374	Cpdm1	insulin level fasting	chrX	20.8	intercross BN/CrjxWBN/KobSlc	11414764
Rn	L13722	101827289	102001374	Sr2	stress response defensive burying - approach	chrX	3.4	intercross WKYxF344	14722724
Hs	NM_000390	83891368	84074423	CGF1	social cognition	chrX	NONE	NONE	9192895

Figure 1. QTL MatchMaker query and input screenshots. (**A**) A section of a Batch Search output page. The results can be downloaded in Microsoft Excel or HTML format. (**B**) Cross Species Search output page.

DISCUSSION

The underlying concept for QTL MatchMaker is a novel approach to gene annotation. This tool meets the challenge of large-scale genomic experiments providing genetic and (patho)physiologic annotation of putative candidate genes for human diseases and prediction of their functions, biological associations and interactions. In an effort to generate cross-reference for experimentally-defined genes and the genetic basis of QTL in rodents and humans, this application integrates functional genomics data (i.e. microarrays gene expression datasets and mutagenesis screen results) with genetic and phenotype/disease traits association data.

In summary, QTL MatchMaker provides a platform for comparative genomics studies of QTL across human, mouse and rat species. Candidate genes can be identified by integrating phenotypic QTL with variation in gene expression. QTL MatchMaker should be a useful resource to address the hitherto unmet need of linking gene expression and genetic determinants of complex traits and disease.

FUTURE DIRECTIONS

QTL MatchMaker tool is a work in progress. It will continue to acquire QTL data through electronic data submission and data mining. In addition, development of ontology annotation based on biomedical parts of ULMS Thesaurus and MeSH trees would permit for keyword searches of the QTL database in the near future. Future improvements will also include graphical representation of the results as well as alignment of syntenic blocks containing QTL for identical traits across species to narrow QTL borders *in silico*.

ACKNOWLEDGEMENTS

We thank Drs Bernice Morrow, Joe Locker, Katalin Susztak and Barbara Birshtein for critical reading and advice. This work was supported by National Institutes of Health grants RO1 DK060043, DK056077 and UO1 DK060995 to E.P.B. Funding to pay the Open Access publication charges for this article was provided by NIH grants R01 DK060043, R01 DK056077, U01 DK060995.

Conflict of interest statement. None declared.

REFERENCES

1. Nadeau,J.H. and Frankel,W.N. (2000) The roads from phenotypic variation to gene discovery: mutagenesis versus QTL. *Nature Genet*., **25**, 381–384.
2. Korstanje,R. and Paigen,B. (2002) From QTL to gene: the harvest begins. *Nature Genet*., **31**, 235–236.
3. Flint,J., Valdar,W., Shifman,S. and Mott,R. (2005) Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Rev. Genet*., **6**, 271–286.
4. Eppig,J., Bult,C., Kadin,J., Richardson,J., Blake,J., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S., Bello,S.M. *et al*. (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res*., **33**, D471–D475.
5. de la Cruz,N., Bromberg,S., Pasko,D., Shimoyama,M., Twigger,S., Chen,J., Chen,C., Fan,C., Foote,C., Gopinath,G. *et al*. (2005) The Rat Genome Database (RGD): developments towards a phenome database. *Nucleic Acids Res*., **33**, D485–D491.
6. Hamosh,A., Scott,A., Amberger,J., Bocchini,C. and McKusick,V. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.
7. Maglott,D., Ostell,J., Pruitt,K. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
8. Kent,J., Sugnet,C., Furey,T., Roskin,K., Pringle,T., Zahler,A. and Haussle,D. (2002) The Human Genome Browser at UCSC. *Genome Res*., **12**, 996–1006.
9. Jansen,R.C. and Nap,J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet*., **17**, 388–391.
10. Peterson,K., Huang,J., Zhu,J., D'Agati,V., Liu,X., Miller,N., Erlander,M., Jackson,M. and Winchester,R. (2004) Characterization of heterogeneity in the molecular pathogenesis of lupus nephritis from transcriptional profiles of laser-captured glomeruli. *J. Clin. Invest*., **113**, 1722–1733.